Developing A Flight Delay Prediction ModelUsing Machine Learning

Project domain : Data science

Team ID:PNT2022TMID36549

TEAM LEAD

Akshaya.S(112719104003)

TEAM MEMBERS

Dhanisha.R.M(112719104009)

Jayasree.M(112719104016)

Pooja.B(112719104027)

ST.PETER'S COLLEGE OF ENGINERING AND TECHNOLOGY

Avadi -600 054

DEVELOPING A FLIGHT DELAY PREDICTION MODEL USING MACHINE LEARNING

Bonafide record of work done by

Akshaya.S (112719104003)

Dhanisha.R.M (112719104009)

Jayasree.M (112719104016)

Pooja.B(112719104027)

Project report submitted in partial fulfillment of the requirements for the degree of

BACHELOR OF ENGINEERING

BRANCH: INFORMATION TECHNOLOGY

Of Anna University

NOVEMBER 2022

CONTENTS

CHAPTER 1
INTRODUCTION

Flight delay is inevitable and it plays an important role in both profits and loss of the airlines. An accurate estimation of flight delay is critical for airlines because the results can be applied to increase customer satisfaction and incomes of airline agencies . We took several independent variables such as humidity, pressure and precipitation for predicting the flight delay. In this project we uses regression techniques of machine learning to predict the flight delay. By predicting the flight delay beforehand, the airport can reduce the cost of waiting.

.

PROJECT OVERVIEW

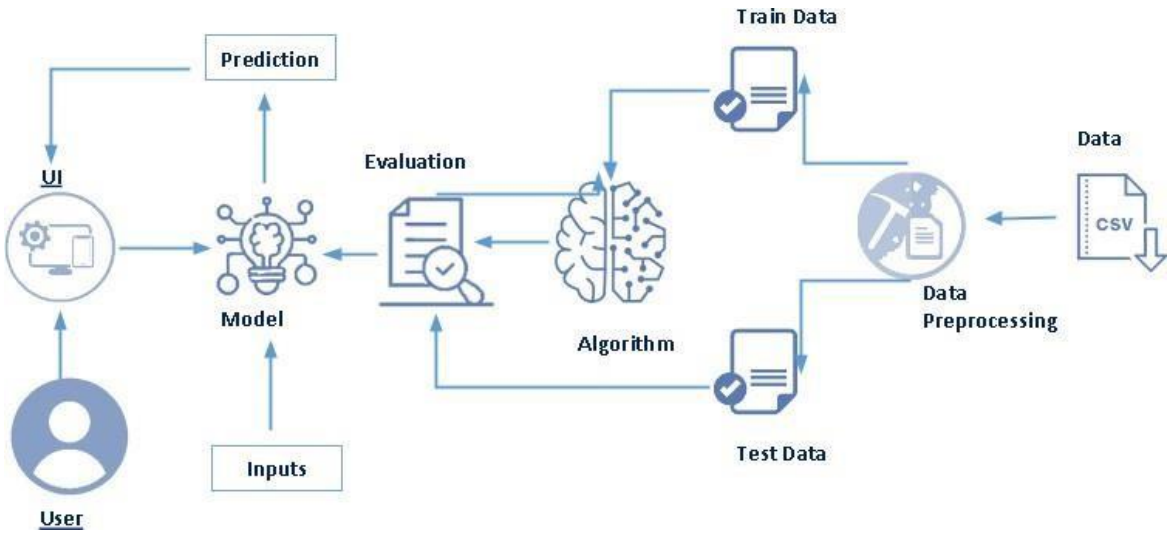


Figure 1.1. Technical Architecture

Flight arrival delays can be predicted using a machine learning algorithm. Rows of feature vectors, such as departure date, delay, travel time between the two airports, and scheduled arrival time, provide the input to our algorithm. The decision tree classifier is then used to determine whether or not the flight arrival will be delayed. When there is more than a 15-minute gap between the scheduled and actual arrival timings, a flight is deemed to be delayed. For various figures of merit, we contrast the decision tree classifierwith logistic regression and a straightforward neural network.

PURPOSE

The main goal of this project is to predict the flight delay using machine learning algorithms. Flight planning is one of the difficulties in the industrial environment because there are many un predictabilities. One such condition is the incidence of delays, which can result from a variety of causes and impose significant expenses on airlines, operators, and passengers. Delays in departure can be brought on by inclement weather, seasonal and holiday demands, airline policies, technical issues with airport infrastructure, baggage handling, and mechanical equipment, and a buildup of delays from earlier flights. Hence Predicting flight delays can improve airline operations and passenger satisfaction, which will result in a positive impact on the economy.

CHAPTER 2

LITERATURE

SURVEY

YEAR	AUTHOR	METHODOLOGY	OBJECTIVES	FINDINGS	PROBLEMS
2017	Gopalakrishnan Balakrishnan	Variables like quarter of the year, year, month, day of month, and day of week have been used to get insights on seasonality. Other variables that were investigated include the origin and destination airports and departure and arrival times.	After the identification of input variables, we collected the relevant data on those variables. Several classification models were then trained and tested. The data collection process data pre-processing in data transformation and description of classification models . After the identification of input variables, we collected the relevant data on those variables. Several classification models were then trained and tested. The data collection process data pre-processing in data transformation and description of classification models .	Variables to be used to study the impact on airline On-Time performance.	For a decision maker, understanding the factors that may impact the outcome of their decision is important.
2020	Maryam Farshchian Yazdi, Seyed Reza Kamel	As the air travels have a signifcant role in economyof agencies and airports, it is nec-essary for themto increasequality of theirservices. One ofthe important modern life challenges of airports andairline agenciesis fight delay.	In this section, we issues to represent a technique in which we tried to solve the problems related to massive data, processing complications lack ofcomputational space,over fitting and existingnoise in data	The model is designed using Machine Learning inTensor flow and is installed on a system of 40 core CPU at frequency of 2.6 hz.	Flight delay has a negative impact on paasengers, Agencies and airports. So, estimation of Flight delay is useful.

2020	Yogitha bose Vital vora	Flight Planning is one of the challenges in industrial world which faces many uncertain conditions. One such condition is delay occurrence, which stems from various factors and imposes considerable costs on airlines,operators, and travelers.	Various methodology can be applied to implement the system that predicts the delay in flight. Decision Tree: As the name suggest the main idea behind decision tree algorithm is to make a tree like structure and get the answers in form oftrue of false. Themodel begins from aroot node and ends on the decision	As discussed, weather condition plays an important role in proper And timely functioning of flights. We propose a flight delay prediction system which focuses mainly on predicting delay of a flight based on the weather situation	As discussed, considering thestandard Taxonomy of the flight delay and its problems, one will contemplate the scope of prediction. Statistical model requires the use of correlation analysis, parametric and non parametric tests,analysis and econometric models.
2020	L. Carvalho A. Sternberg	The mainproblems relatedto flight delay prediction are identified and organized in a taxonomy. It includes scopes, models, and ways of handling flight delay prediction.	The flight delay prediction problem may be modeled in many ways, depending on the objectives of the research. Methods were divided into five groups they are Statistical analysis, Probabilistic, Network, Machine learning, Operational.	Root delay propegation Researchers create prediction models to tackle root delay, predicting when and where a delay will occur and what are its reasons and sources.	Problem is the core feature in Domain taxonomy. As seen there are three major concerns regarding the flight delay prediction Problem: delay propagation, root delay and cancellation.

2021	Viran Raj Satyam Singh	We are proposing machine learning algorithms like Linear regression Techniques. The aim of this research work isto predict Flight Delay.	A. Using Linear Regression:Sincelogistic regression is appropriate for categorical values, and we expect to predict the potential delayed time, which is a numerical valuable, it makes more sense to apply Linear Regression for our model. B. Using Logistic Regression: Actual Arrival time - Expected Arrival Time + (Actual Departure time - Expected Departure .	A.Logistics Regression Model B.Linear Regression Model C. Initial Data Exploring D. Dimensionality Reduction.	No data mining projects could be finished without thoroughly understand the data first.We renamed the original data column names and validated the nulls, however with a little different approach. After data cleaningwe start the first process of exploring our data if there were any patterns within the independent variables.

Existing System

Airlines, airports, and passengers would all benefit from a more accurate flight delay prediction model. Currently, models used by airlines to predict flight delays are based on historical data and do not take into account real-time data such as weather conditions. This can lead to delays and cancellations, as well as increased cost.

Reference Link

https://ieeexplore.ieee.org/document/8903554
https://ieeexplore.ieee.org/document/9512525
https://www.sciencedirect.com/science/article/abs/pii/S1270963821003321
https://ieeexplore.ieee.org/document/9129110

CHAPTER 3

IDEATION & PROPOSED SOLUTION

EMPATHY MAP CANVAS

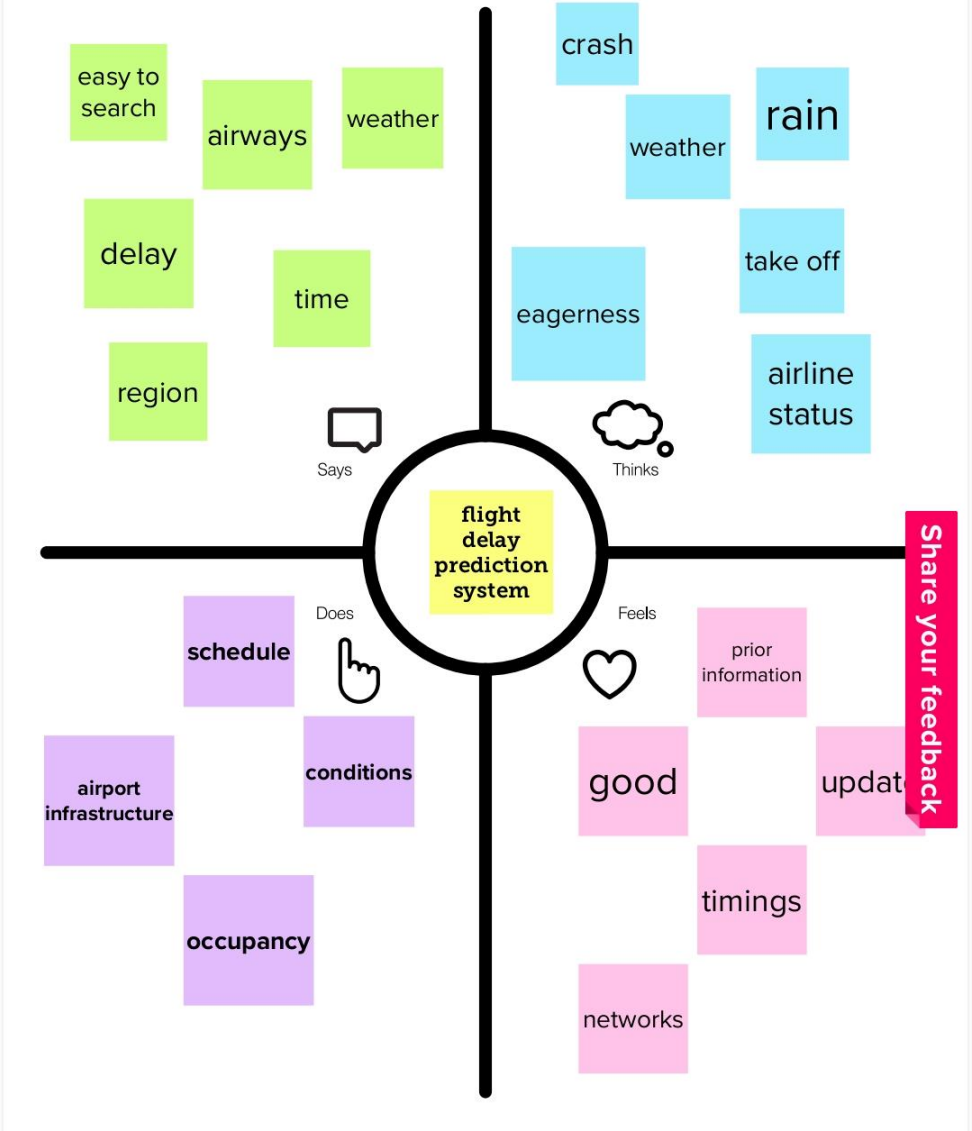
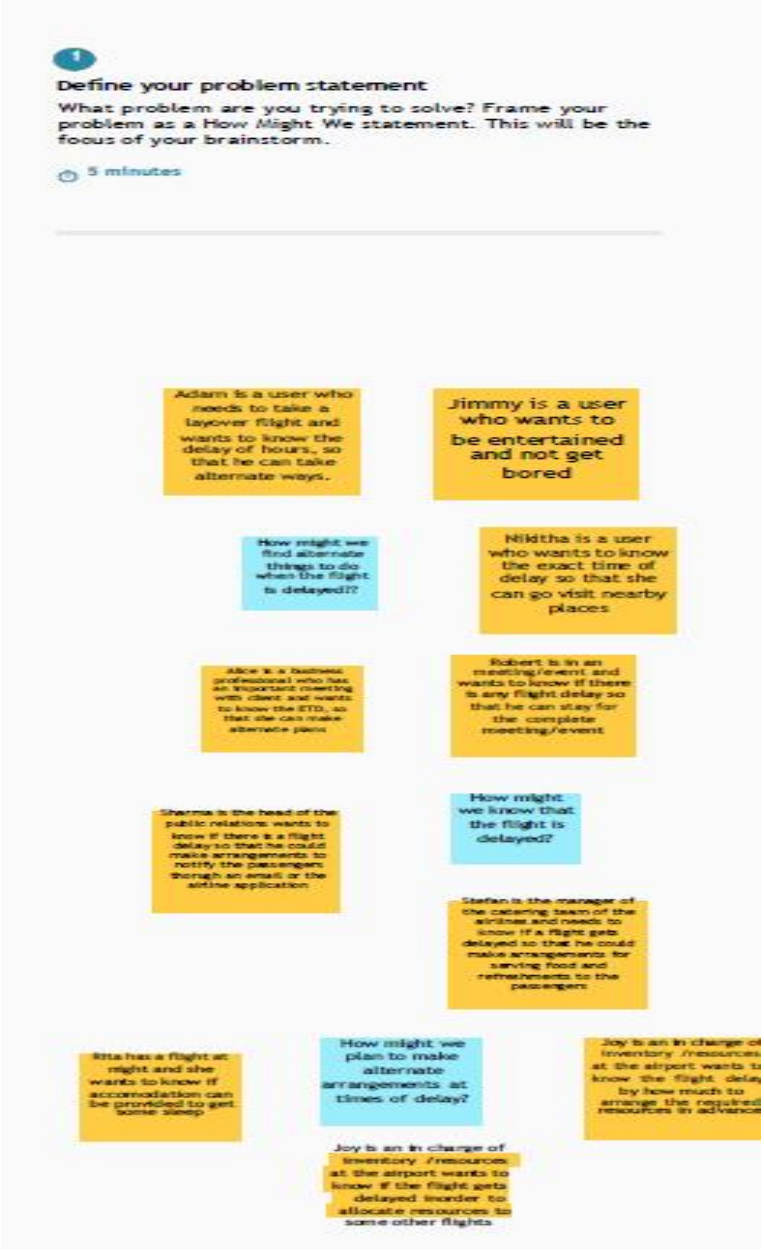


Figure 3.1. Empathy Map

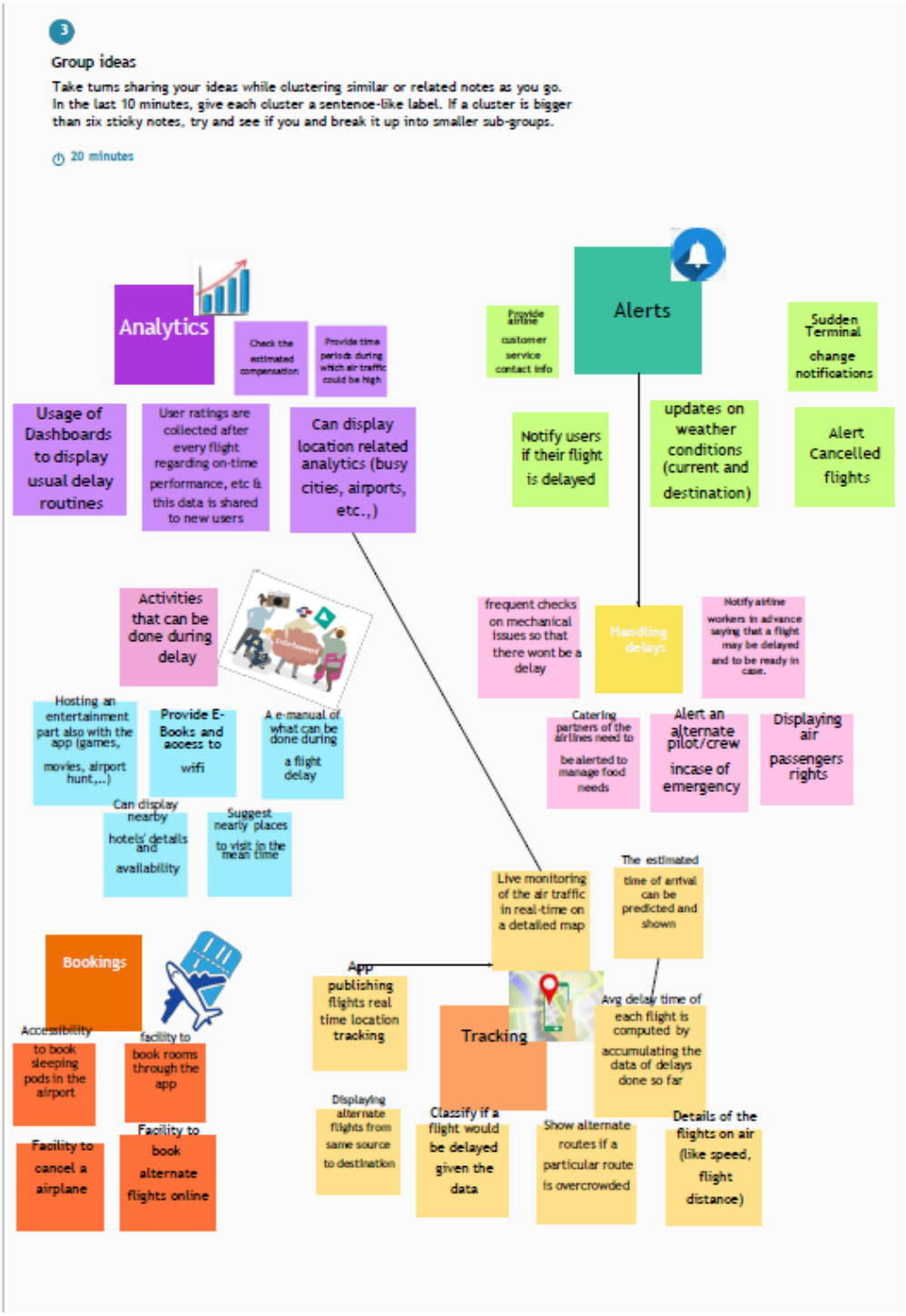
IDEATION & BRAINSTORMING

Step 1 - Team Gathering, Collaboration and Selecting the Problem Statement

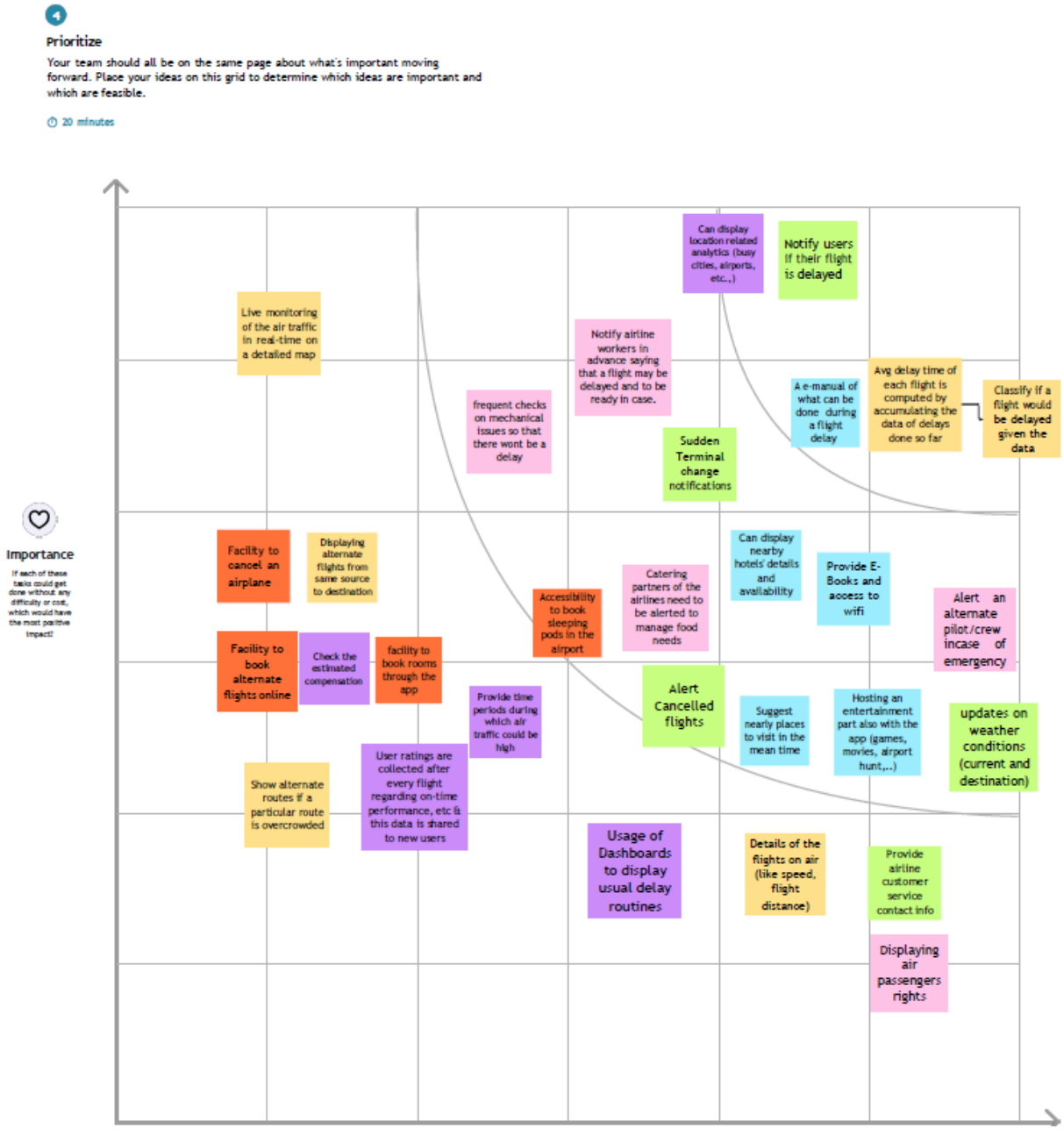


Step 2 - Brainstorm

Step 3 – Grouping ideas



Step 4 - Idea Prioritization

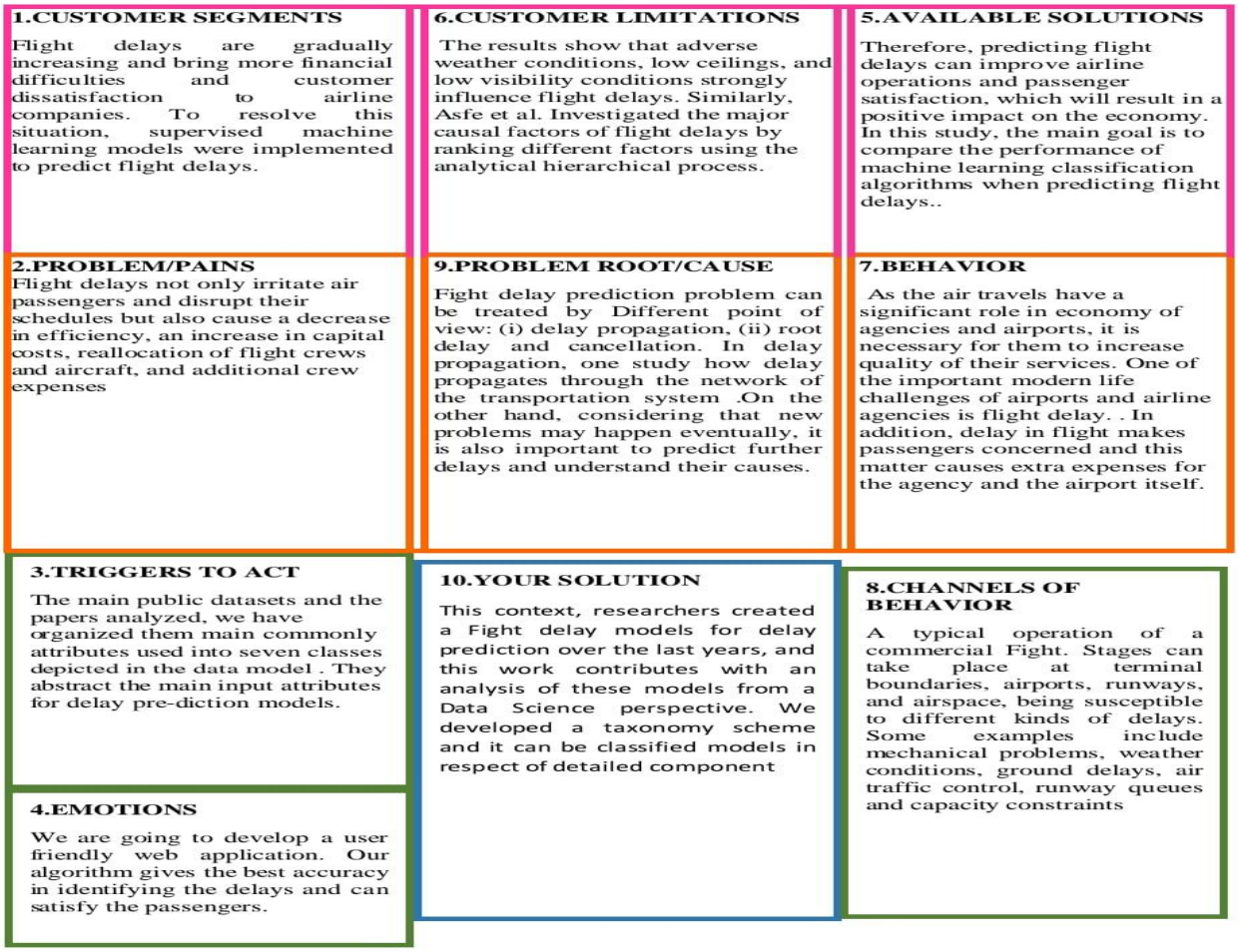


3.3. PROPOSED SOLUTION

S.No.	Parameter	Description
1	Problem Statement (Problem to be solved)	Flight arrival delays can be predicted using a machine learning algorithm. Rows of feature vectors, such as departure date, delay, travel time between the two airports, and scheduled arrival time, provide the input to our algorithm. The decision tree classifier is then used to determine whether or not the flight arrival will be delayed. When there is more than a 15-minute gap between the scheduled and actual arrival timings, a flight is deemed to be delayed. Additionally, for various figures of merit, we contrast the decision tree classifier with logistic regression and a straightforward neural network.
2	Idea / Solution description	Using ML Algorithms to predict the delay in flight arrival, informing them to the customers using a Mobile Application or a Web Application. We are developing a software that will allow passengers who use airplanes to foresee flight delays. They may effectively plan their travelusing this application, which will help them save time. Thetool will have an intuitive user interface. To estimate delays and execute the most effective and efficient methods in the tool, we will use a variety of machine learning algorithms.
3	Novelty / Uniqueness	1. Building a full-fledged application in which the customers can track whether the flights will be delayed or not. 2. Combining the results of one or more ML models using the techniques of ensembling
4	Social Impact / Customer Satisfaction	Flight delays not only anger and disturb air travelers' plans,but they also reduce productivity, raise capital costs, reallocate flight crews and aircraft, and add to crew costs. Higher operating costs for airline firms are

		unavoidable as flight delays necessitate the consumption of more labor, capital, and other necessary inputs. Flight delays could make the transportation system less effective and have a negative impact on how an airport is planned. Delayed flights subject airlines to penalties, fines,and additional expenses.
5	Business Model (Revenue Model)	The cost of airline tickets and flight delays are now uncertain. Even for the same airplane and seat class, ticket costs are dynamic and frequently change. To increase their revenue, airline firms use a variety of algorithms to adjust the prices dynamically. These modelsare not accessible to the general public due to the intensecompetition among airline operators. Additionally, the flight is delayed due to a number of micro and macro causes. The air route status, the prior flight's delay, airplane capacity, air traffic management, airline properties, etc. are the main elements that have an impact on airlines. To save "Time and Money," it is necessary to forecast airline flight delays andticket costs.
6	Scalability of theSolution	The proposed system can be scaled up to take actions – book another flight for passengers or if a particular flight is getting delayed often, the same can be examined by memorizing the outputs of this system. This can be scaledup to predict the delay of flights in every airport.

3.4. PROBLEM SOLUTION FIT



CHAPTER 4

REQUIREMENT

ANALYSIS

FUNCTIONAL REQUIREMENT

Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Registered User -Login	Login through password(Form) Login through Gmail Login through LinkedIn
FR-4	Verify the link provided by the user	User inputs the link to be verified
FR-5	Display the result	If the site link is a prediction site,user must be aware and read the precautions displayed If the site link is legit,exit the application
FR-6	Share Queries	If any doubts,send query Read FAQs

NON-FUNCTIONAL REQUIREMENTS

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	Knowing when the flight will be delayed enables improved operational planning at the airport of destination based on anticipated flight delay at origin.
NFR-2	**Security**	It is highly secure and the passengers who log in to the application will be able to view the status
NFR-3	**Reliability**	As we train with more data the model will be reliable.
NFR-4	**Performance**	• This was done statistically, and the delay time was thought to be shorter. • Using variables that occur close to the destination's arrival time, it has projected the delay at the destination.
NFR-5	**Availability**	The mobile application must be accessible to users in India 99.98% of the time each month between EST and IST business hours.
NFR-6	**Scalability**	• The main problem for airlines and travellers is flight delay. • According to the flight schedule, the anticipated arrival delay considers both flight information and the weather at the airports of origin and destination.

CHAPTER 5

PROJECT DESIGN

DATA FLOW DIAGRAMS

A Data Flow Diagram (DFD) is a traditional visual representation of the information flowswithin a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changesthe information, and where data is stored.

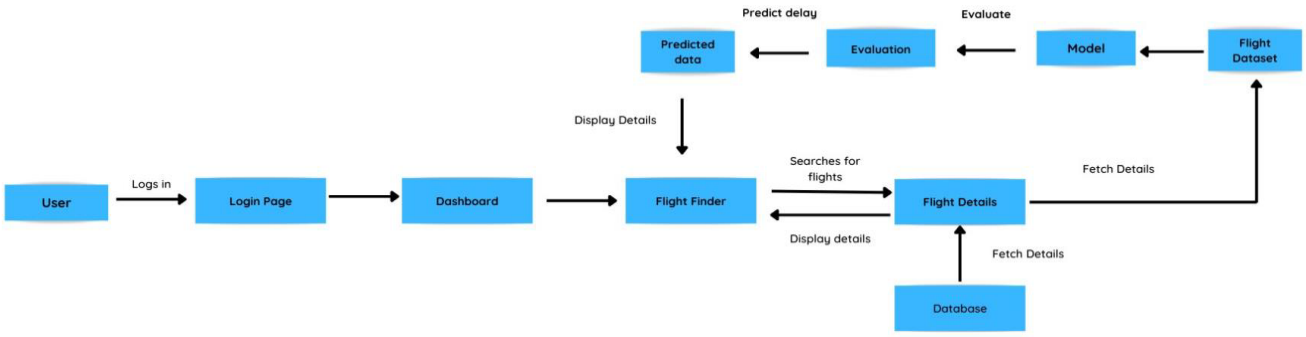
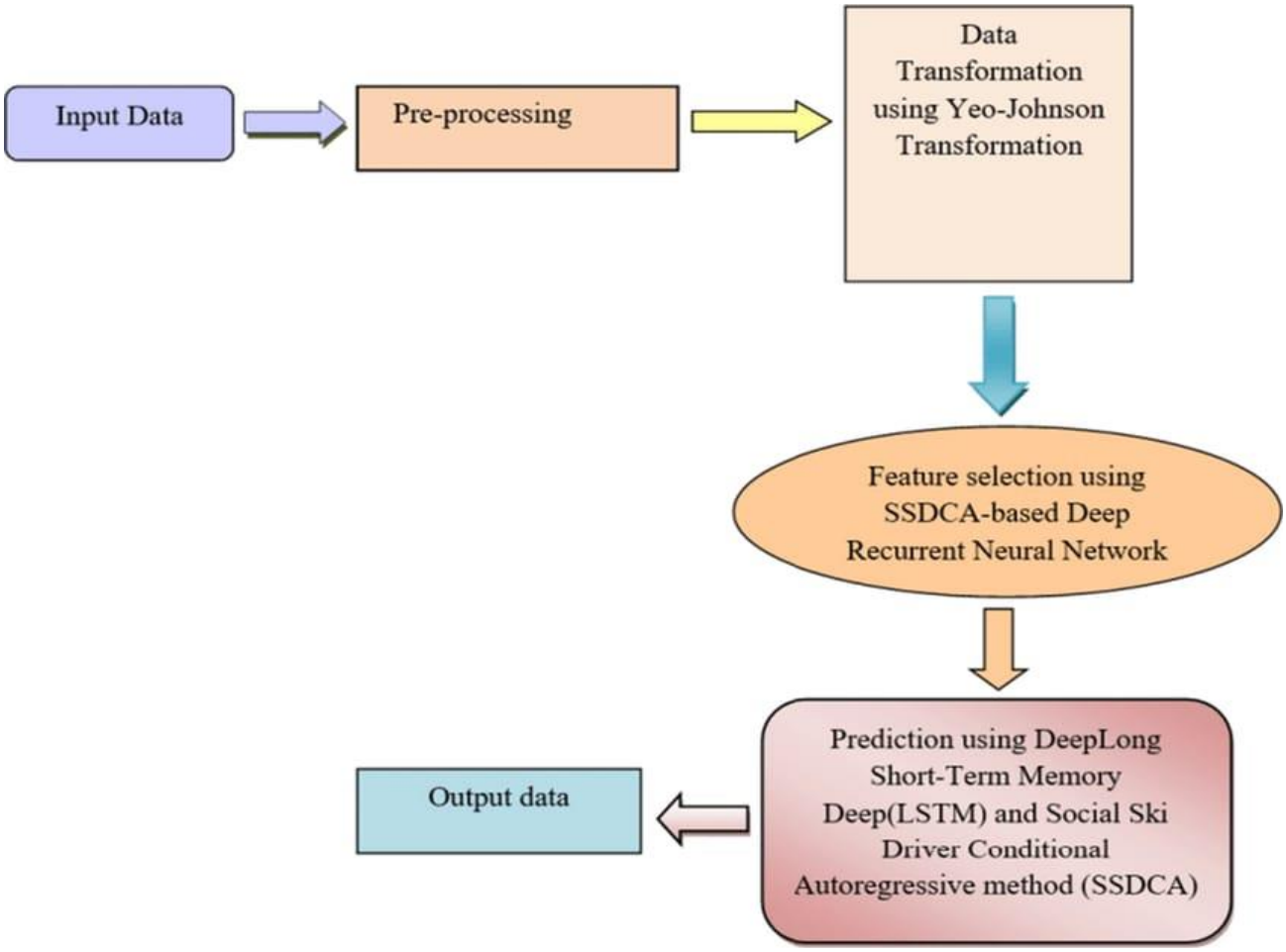


Figure 5.1. Data flow diagram

SOLUTION & TECHNICAL ARCHITECTURE



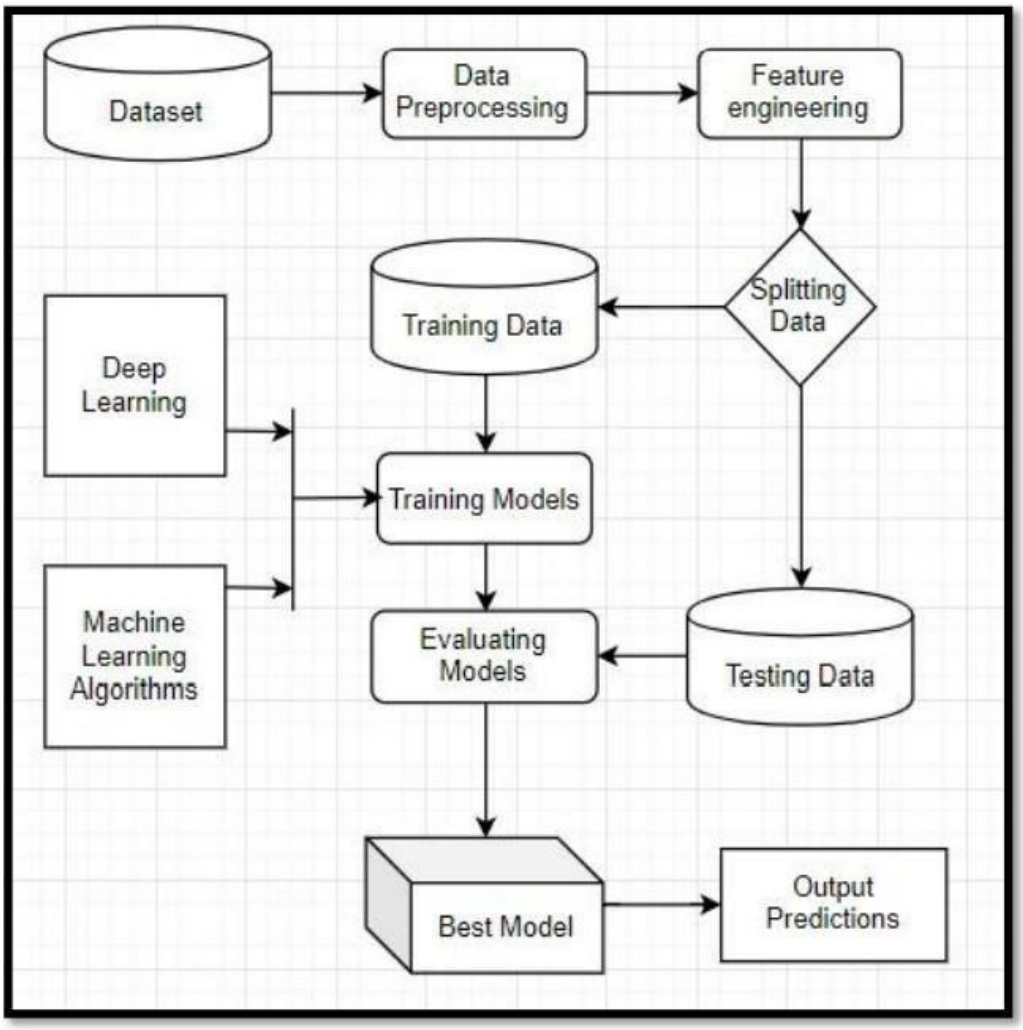


Figure 5.2. Solution Architecture

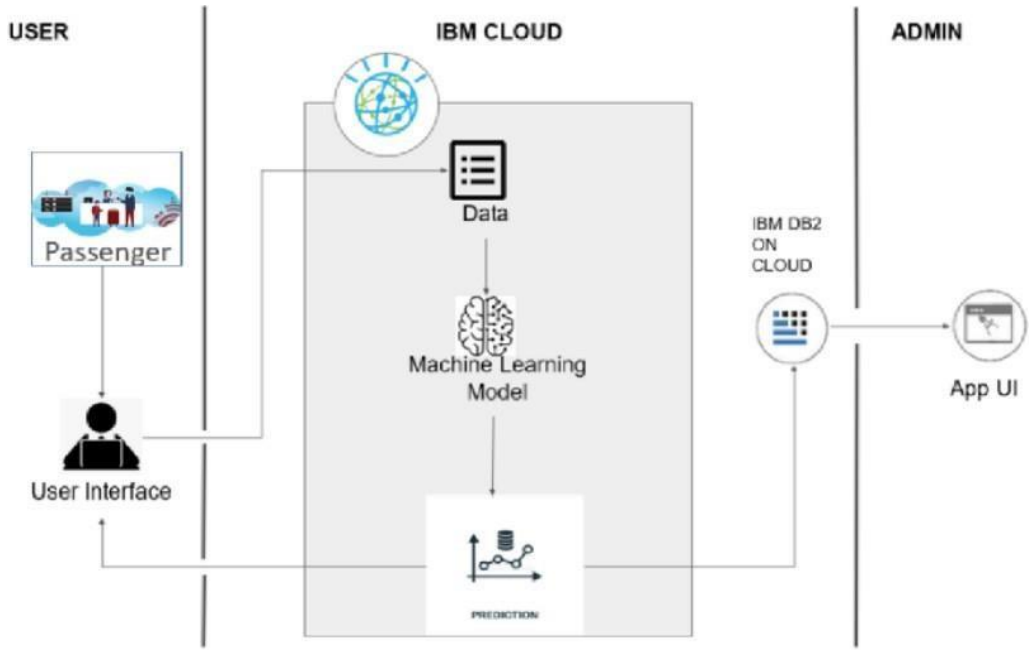


Figure 5.3. Technology Stack

Components & Technologies

S.No	Component	Description	Technology
1.	User Interface	Web Application to interact with the user.	Flask
2.	Login/Sign up	Login/ Sign up – The user can enter the details and get them validated	Python
3.	Database	The Database to store the login details of the user	MySQL
4.	Cloud Database	The database to keep track of the flight details from the travel agency, input to the Machine Learning Model	Firebase
5.	Machine Learning Model	To Predict whether the flight will get delayed or not.	SVM, KNN Classifier, Logistic Regression, Decision Trees
6.	Deep Learning Model	To Predict whether the flight will get delayed or not	Fully Connected Neural Networks
7.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud Local Server Configuration: Cloud Server Configuration :	IBM Cloud

Application Characteristics

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	Web application – Flask ML – Sklearn, Tensorflow, Keras API	Deep Learning, Python
2.	Security Implementations	The data is secured that it is encrypted in IBM cloud	AES (256-bit)
3.	Scalable Architecture	Can be scaled upto many airports, many users with more training	Firebase
4.	Availability	The status will be updated frequently	IBM Cloud
5.	Performance	Can make as many number of requests per second to get the prediction	IBM Cloud

User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail	I can register & access the dashboard with Gmail Login	Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password	I can login & access the dashboard	High	Sprint-1
	Core	USN-6	As a user, I can enter my flight details	I can feed the inputs to the system	High	Sprint-2
		USN-7	As a user, I can look at the flight details	I can see whether my flight is getting delayed or not	High	Sprint-3

CHAPTER 6

PROJECT PLANNING & SCHEDULING

SPRINT PLANNING & ESTIMATION

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data Engineering	USN-1	Data Collection, Data Preprocessing and Feature Extraction	4	High	Sujith Kumar M A
Sprint-1	Machine Learning Prediction Model	USN-2	Building a Machine Model for Flight Delay Prediction, Testing with different metrics.	4	High	Anusha Devi R
Sprint-2	Flask Web Page	USN-3	Building Home Page and Prediction Page.	4	Low	Nandish Chandrasekar
Sprint-1	Integration.	USN-4	Integrating the flask pages with the ML Model and IBM Cloud Deployment	4	Medium	Tamilselvan M

SPRINT DELIVERY SCHEDULE

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	4	6 Days	24 Oct 2022	29 Oct 2022	4	29 Oct 2022
Sprint-2	4	6 Days	31 Oct 2022	05 Nov 2022	4	05 Nov 2022
Sprint-3	4	6 Days	07 Nov 2022	12 Nov 2022	4	12 Nov 2022
Sprint-4	4	6 Days	14 Nov 2022	19 Nov 2022	4	19 Nov 2022

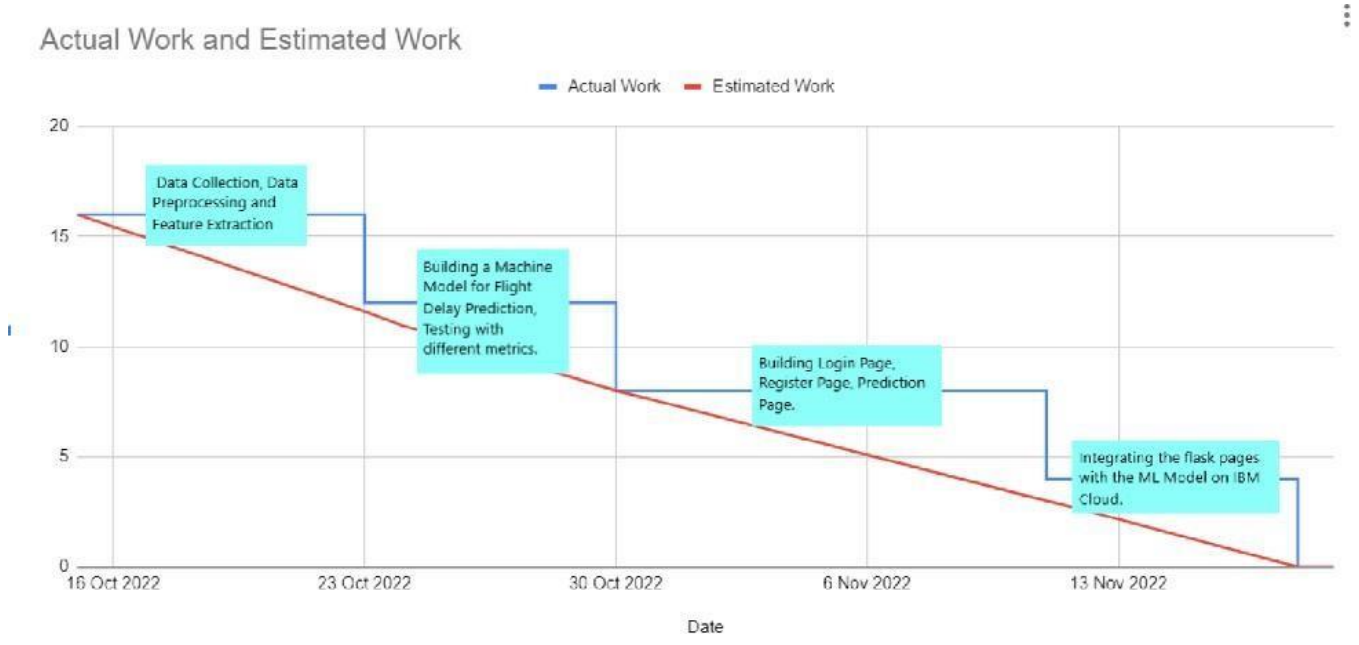


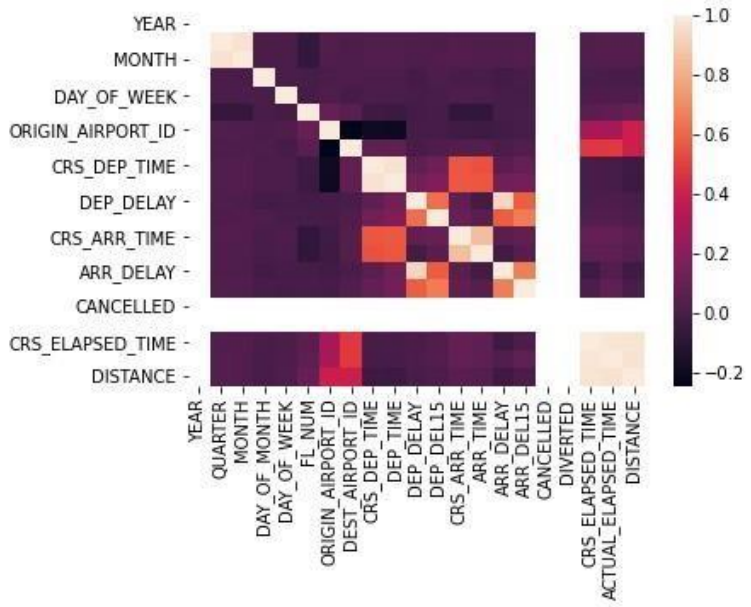
Figure 6.1 - Burndown Chart

CHAPTER 7

CODING AND SOLUTIONING

FEATURE 1 - CORRELATION BETWEEN THE VARIABLES IN THE DATASET

```
In [19]: sns.heatmap(data.corr())
Out[19]: <AxesSubplot:>
```



This will help us to find out the correlation between the variables in the dataset whichwould help us to find out the columns that are unnecessary and hence to be dropped.

FEATURE 2 - ONE HOT ENCODING

```
In [39]: data=pd.get_dummies(data,columns=['ORIGIN','DEST'])

In [40]: data['ARR_DEL15'].value_counts()
Out[40]: 0.0    9668
         1.0    1375
         Name: ARR_DEL15, dtype: int64

In [41]: data.tail()
Out[41]:
```

	FL_NUM	MONTH	DAY_OF_MONTH	DAY_OF_WEEK	CRS_ARR_TIME	DEP_DEL15	ARR_DEL15	ORIGIN_0	ORIGIN_1	ORIGIN_2	ORIGIN_3	ORIGIN_4
11226	1715	12	30	5	12	0.0	0.0	0	1	0	0	0
11227	1770	12	30	5	20	1.0	0.0	0	0	0	0	1
11228	1823	12	30	5	22	0.0	0.0	0	1	0	0	0
11229	1901	12	30	5	18	0.0	0.0	1	0	0	0	0
11230	2005	12	30	5	9	0.0	0.0	1	0	0	0	0

The cities in both Origin and Destination are one-hot encoded using the above code

CHAPTER 8

TESTING

TEST

User No	Flight No	Month	Day of month	Day of week	Origin	Destination	Scheduled Departure Time	Scheduled Arrival Time	Actual Departure Time	Actual Inputs
1	1232	1	1	1	ATL	MSP	1905	2305	1945	Delayed
2	1399	1	1	1	ATL	SEA	1805	2410	1855	Delayed
3	2351	1	2	3	ATL	DTW	1305	2305	1305	Not Delayed
4	2637	2	1	3	DTW	ATL	1500	2410	1505	Not Delayed

USER ACCEPTANCE TESTING

This report shows the number of test cases that have passed and failed

User No	Flight No	Month	Day Of Month	Day Of Week	Origin	Destin -ation	Scheduled Departure Time	Scheduled Arrival Time	Actual Departure Time	Actual Output	Predict -ed Output	Correct-ness
1	1232	1	1	1	ATL	MSP	1905	2305	1945	Delayed	Delayed	Correct
2	1399	1	1	1	ATL	SEA	1805	2410	1855	Delayed	Delayed	Correct
3	2351	1	2	3	ATL	DTW	1305	2305	1305	Not Delayed	Not Delayed	Correct
4	2637	2	1	3	DTW	ATL	1500	2410	1505	Not Delayed	Not Delayed	Correct

CHAPTER 9
RESULTS

PERFORMANCE METRICS

Training Accuracy

MODEL EVALUATION

```
acc=accuracy_score(predicted,y_test)
```

```
acc
```

```
0.8791308284291535
```

Confusion Matrix

```
from sklearn.metrics import confusion_matrix
confusion_matrix(predicted, y_test)
```

```
array([[1825,  129],
       [ 138,  117]], dtype=int64)
```

Classification Model

```
from sklearn.metrics import classification_report
print(classification_report(predicted, y_test, labels=[1, 2, 3]))
```

```
              precision    recall  f1-score   support

           1       0.48      0.46      0.47       255
           2       0.00      0.00      0.00         0
           3       0.00      0.00      0.00         0

   micro avg       0.48      0.46      0.47       255
   macro avg       0.16      0.15      0.16       255
weighted avg       0.48      0.46      0.47       255
```

CHAPTER 10
ADVANTAGES AND DISADVANTAGES

Advantages

- Customers are happy
- The available flights are easily identified
- Prior information will be sent if in case the flight is delayed
- The current status of the flight can be tracked

Disadvantages

- Wrong prediction due to noise of input data
- If the prediction is wrong, then there will be extra expenses for the agencies, passengers and airport
- Passengers with medical emergencies gets affected

CHAPTER 11

CONCLUSION

In this project, we use flight data, weather, and demand data to predict flight departure delay. In the end, our model correctly predicts the delayed and non-delayed flights correctly. As a result, there can be additional features related to the causes of flight delay that are not yet discovered using our existing data sources.

CHAPTER 12
FUTURE SCOPE

Based on data analysis from the year 2008, this project. There is a sizable dataset accessible from 1987 to 2008, but managing a larger dataset necessitates extensive preprocessing and purification of the data Therefore, adding a larger dataset is a part of this project's future effort. Preprocessing a bigger dataset can be done in a variety of methods, such as establishing a Spark cluster on a computer or using cloud services likeAWS and Azure. Now that deep learning has advanced, we can employ neural networksalgorithms to analyze aviation and meteorological data. Neural networks employ a formof pattern matching.

The project's focus is primarily on flight and weather data for India, but we can alsoinclude data from other nations like China, the United States, and Russia. We can broaden the project's scope by including flight information from international flights ratherthan just domestic flights.

CHAPTER 13

APPENDIX

Source code

Flask Application

```
from flask import Flask, request, render_template
import numpy as np
import pandas as pd
import pickle
import os
model = pickle.load(open('flight_new.pk1','rb'))
app = Flask(     name         )
@app.route('/')
def home():
    return render_template("mainpage.html")


@app.route('/prediction',methods=['GET','POST'])def
predict():
    name = request.form['fname']
    month = request.form['month']
    dayofmonth = request.form['daymonth']
    dayofweek = request.form['dayweek']
    origin = request.form['origin']
    if(origin == "msp"):
        origin1, origin2, origin3, origin4, origin5 = 0,0,0,0,1
    if(origin == "dtw"):
        origin1, origin2, origin3, origin4, origin5 = 1,0,0,0,0
    if(origin == "jfk"):
        origin1, origin2, origin3, origin4, origin5 = 0,0,1,0,0,
    if(origin == "sea"):
        origin1, origin2, origin3, origin4, origin5 = 0,1,0,0,0
```

```
    if(origin == "atl"):
        origin1, origin2, origin3, origin4, origin5 = 0,0,0,1,0
    destination = request.form['destination']
    if(destination == "msp"): destination1,destination2,destination3,destination4,destination5 =
        0,0,0,0,1
    if(destination == "dtw"): destination1,destination2,destination3,destination4,destination5 =
        1,0,0,0,0

    if(destination == "jfk"): destination1,destination2,destination3,destination4,destination5 =
        0,0,1,0,0
    if(destination == "sea"): destination1,destination2,destination3,destination4,destination5 =
        0,1,0,0,0
    if(destination == "atl"): destination1,destination2,destination3,destination4,destination5 =
        0,0,0,1,0
    dept = request.form['sdeparttime']
    arrtime = request.form['sarrivaltime']
    actdept = request.form['adeparttime']
    dept15 = int(dept)-int(actdept)
    total =
[[name,month,dayofmonth,dayofweek,arrtime,dept15,origin1,origin2,origin3,origin4,origi
n5,destination1,destination2,destination3,destination4,destination5]]
    y_pred = model.predict(total)
    print(y_pred)
    if(y_pred == [0.]):
        ans = "The Flight will be on time"
    else:
        ans = "The Flight will be delayed"
    return render_template("index.html",data = ans)
app.run(debug=True)
```

Github Link:

https://github.com/IBM-EPBL/IBM-Project-36932-1660298968

Video Demo Link:

https://drive.google.com/drive/u/0/folders/1DX8Jpt0-uB-VRfUFkrVJAubv8VK0lWHc